Application of Machine Learning Models for Movie Box Office Prediction and Data Exploration

Victoria Anorve, Joel Pepera, Benya Piyathamsoontorn, Radha Yadavilli, & Dan Zhou

January 10th, 2020

Keywords

Box Office Predictions, Python, SQL, Machine Learning, Model Training, Real-time prediction APP

Forecasting box office revenue using only information available pre-release involves understanding various dynamics impacting consumer preference, and traditional methodology has shown only modest performance [1]. Some research groups have implemented a more comprehensive data science framework to combine survey data [2] to improve accuracy of forecasting. Our project will apply a variety of machine learning models to predict film performance based only on information available before a film's release.

Introduction:

Movie box office revenue prediction is an important issue in the film industry. The application of machine learning models for planning and predicting revenues has become somewhat common, aiming to improve the movie production and selection process. Generally, prediction is based on using data collected from the multiple social media, internet social sources including YouTube, Twitter, Box Office Mojo, IMDb, The Numbers, etc. [3]

Additionally, the movie consumer behavior continues to change as home-streaming applications have dominated the consumer market. For instance, Netflix has altered the way people view movies because of its growing library of affordable and convenient products.

The success or failure rate of a movie can depend on several features: release period, investment budget, actor/actress selection, movie rating, production studio, director and so on [4]. Our capstone project aims to develop the best model based on data mining techniques that will help in predicting the success of movie box office sales for the purpose of reducing the risk of movie production costs.

Throughout this project, we will explore the following questions:

- How much variation is there in box office performance by movie?
- What attributes show the strongest correlation to sales?
- What level of prediction accuracy is possible with the data we have?
- How much revenue will a particular movie gross in sales?

We hypothesized that by using information available before the movies' release, we can significantly improve baseline predictions.

Data Preparation: ingestion, wrangling, computation, visualization

Our dataset was assembled from the following websites: Box Office Mojo, Internet Movie Database (IMDb), and The Numbers. We first imported data from multiple files available through IMDB.com, with each dataset contained in a zipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. We uploaded these files to a PostgreSQL database hosted on Amazon's AWS RDS, to which we connected through pgAdmin and Python/Jupyter. For the other the Box Office Mojo and Numbers sources, we used a web-scraping technique to collect the URL's of 10 web pages with the *requests* Python package. We used *beautifulsoup* to parse the data, which was also stored in our PostgreSQL database.

A snapshot of our web-scrapping process:

```
In [1]:
# Import package
import pandas as pd
import requests
from bs4 import BeautifulSoup

# Specify url: url
res = []
url = 'https://www.the-numbers.com/movie/budgets/all'
urllist = [url]
# urllist = []
numofiter = 59

In [2]:
def urlAppend(urllist,url):
    return urllist.append(url)

def genUrl(url,b):
    return url+"/"+str(b)

In [3]:
j=1
for i in range(1,numofiter):
    j+=100
    s=genUrl(url,j)
    urlAppend(urllist,s)
    print(s)

https://www.the-numbers.com/movie/budgets/all/101
https://www.the-numbers.com/movie/budgets/all/201
https://www.the-numbers.com/movie/budgets/all/301
https://www.the-numbers.com/movie/budgets/all/401
https://www.the-numbers.com/movie/budgets/all/501
https://www.the-numbers.com/movie/budgets/all/601
https://www.the-numbers.com/movie/budgets/all/701
https://www.the-numbers.com/movie/budgets/all/801
https://www.the-numbers.com/movie/budgets/all/901
https://www.the-numbers.com/movie/budgets/all/1001
```

After filtering the IMBD data (which contained most of the movie attributes) for titles released in the United States after 1999, we were left with 73,626 rows. We ran into some issues in joining this data to that from The Numbers and BoxOfficeMojo (which added our sales target information and some additional attributes like keywords and ratings) due to a mismatch of title availability across sources and inconsistency in the way title names were listed. To help mitigate this entity resolution challenge, we applied a 'slugify' function to remove spaces, case, and punctuation from title names, which modestly improved our match rate. Even so, our final table for modeling only included 2,788 rows.

We went through multiple iterations to determine how to represent genres, movie description keywords, actors/actresses, directors, and others associated with each movie. These attributes are inherently many-to-one relationships, and the association level with one film may be different than that with another. For example, even if Title A shares three actors/actresses with

Title B, the lead actor may be different. Similarly, some films may be listed with Action as the primary genre and Comedy as secondary, while others may list Comedy as primary and Action as secondary or tertiary. Will the order of listing add predictive power to our models?

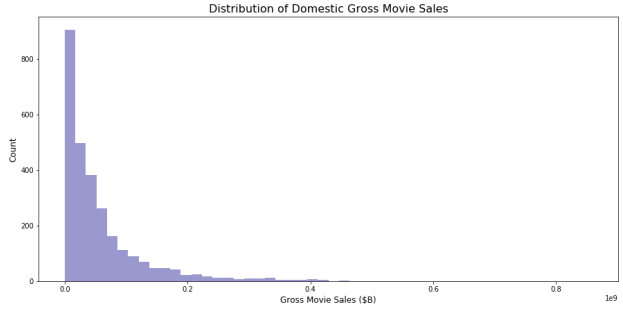
Initially, we tried preserving the order in which these attributes were listed, so we included fields for the first actor and genre listed, the second actor and genre, etc. However, we ultimately decided against this because: a) it divided the representation of an individual actor/keyword/genre across multiple columns, making a sparse feature even more sparse, and b) it made the interpretation of features more difficult (e.g. we couldn't simply determine the impact of having a particular actor in a film). Thus, we opted to simply one-hot encode each actor, actress, director, keyword, and genre, and we limited the dimensionality by judgmentally selecting a minimum number of movies for which the feature needed to be present in order to include that feature. Additionally, we experimented with PCA and various clustering techniques to reduce the dimensionality of genre features (e.g. encoding 'Action-Adventure' and 'Romantic Comedy' instead of treating them separately), but this failed to increase model performance.

Exploratory Data Analysis

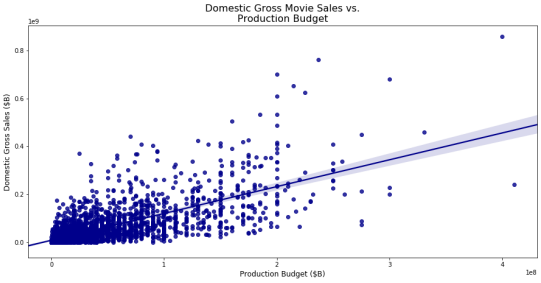
We examined the sales distribution across many key variables in our dataset, including:

> Release Date, Production Budget, Movie Rating, Studio, Movie Runtime, genre, Keyword Descriptions, Directors, Actors/Actress

Our visualization process involved using Python libraries such as *matplotlib*, and *seaborn* to generate histograms, boxplots, and scatterplots. We found movie sales had a high correlation with production budget, movie runtime, release time, and genres. However, the gross of the movie sale has low correlation with most actors and actresses.



Analysis: This graph looks at the distribution of domestic sales. The data has a right skew. The count of movies decreases as the movie sales increase which suggest that the amount of movies that have high domestic sales are not the norm. Our project wants to examine which features are associated with these high movie sales.



Analysis: -The regression model shows the relationship between production budget and sales. It is the strongest predictor. We can clearly see a positive linear correlation between the two variables. Although deviations exist as indicated by the points scattered across the linear line, the relationship is still strong.

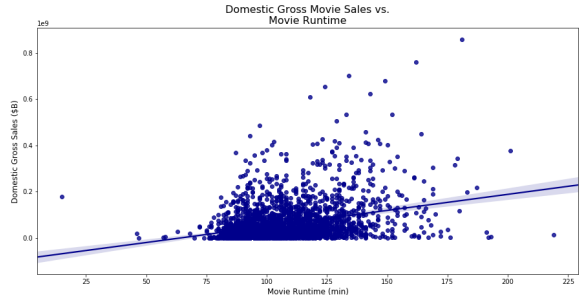
OLS Regression Results

Dep. Variable:	domesticgross	**R-squared:**	0.496
Model:	OLS	**Adj. R-squared:**	0.495
Method:	Least Squares	**F-statistic:**	2930.
Date:	Sat, 04 Jan 2020	**Prob (F-statistic):**	0.00
Time:	00:15:27	**Log-Likelihood:**	-57526.
No. Observations:	2984	**AIC:**	1.151e+05
Df Residuals:	2982	**BIC:**	1.151e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	7.169e+06	1.4e+06	5.121	0.000	4.42e+06	9.91e+06
productionbudget	1.1271	0.021	54.129	0.000	1.086	1.168

Omnibus:	1552.797	**Durbin-Watson:**	1.640
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20805.348
Skew:	2.152	**Prob(JB):**	0.00
Kurtosis:	15.199	**Cond. No.**	9.01e+07

Analysis: - The strong/moderate relationship between sales and production budget is indicated by the R-squared value, 0.495. We also look at the p-value, which lies at 0.0 indicating that there is a statistically significant relationship between the two variables. Lastly, the standard error value, 0.021 is small. This small value indicates that the variation of random variables has a small spread.



Analysis: - This graph looks at the relationship between movie runtime and our output. We see that the relationship is positively correlated but it does not appear to be as strong as the relationship between movie production and sales. We also see that many of the points deviate from the line indicating that the relationship is not as strong.

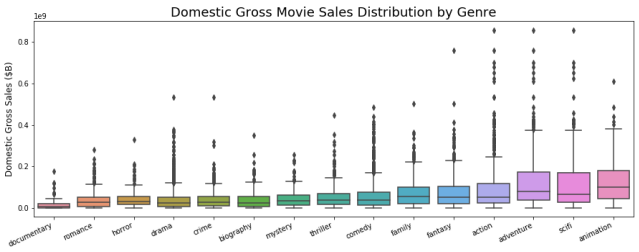
OLS Regression Results

Dep. Variable:	domesticgross	**R-squared:**	0.101
Model:	OLS	**Adj. R-squared:**	0.101
Method:	Least Squares	**F-statistic:**	333.4
Date:	Sat, 04 Jan 2020	**Prob (F-statistic):**	1.12e-70
Time:	00:29:30	**Log-Likelihood:**	-58118.
No. Observations:	2970	**AIC:**	1.162e+05
Df Residuals:	2968	**BIC:**	1.163e+05
Df Model:	1		
Covariance Type:	nonrobust		

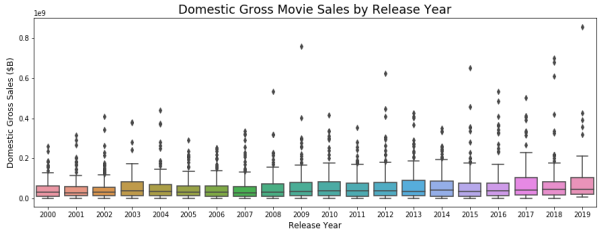
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-9.233e+07	8.34e+06	-11.064	0.000	-1.09e+08	-7.6e+07
runtimeminutes	1.404e+06	7.69e+04	18.259	0.000	1.25e+06	1.55e+06

Omnibus:	1925.718	**Durbin-Watson:**	1.492
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27035.018
Skew:	2.900	**Prob(JB):**	0.00
Kurtosis:	16.595	**Cond. No.**	647.

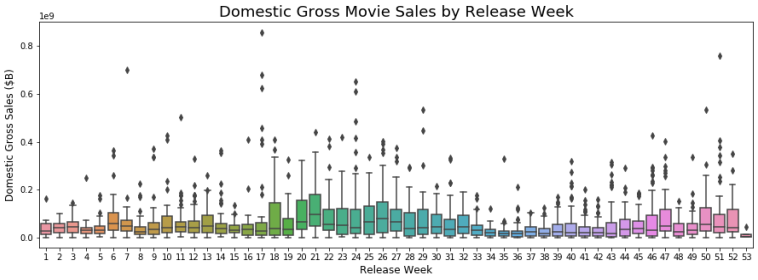
Analysis: The summary statistics show that the relationship between runtime minutes and sales is not as strong as the relationship between production budget and sales. We have an r-squared value of 0.101 and although our p-value is statistically significant, the spread of random variables is large, which lies at 7.69 to the fourth power.



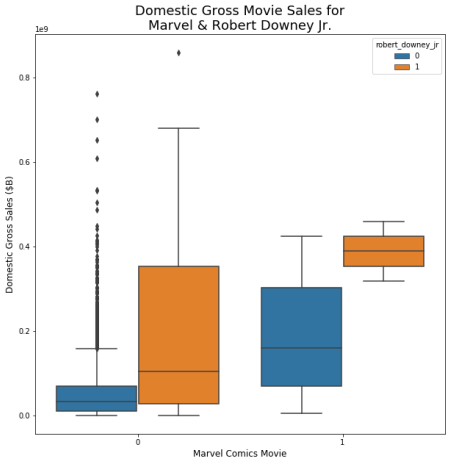
Analysis: This distribution shows the genres such as action, adventure and sci-fi have garnered the highest amount of domestic gross sales. This makes sense because the fast and furious franchise, the avengers franchise and avatar were top grossers in the movie industry.



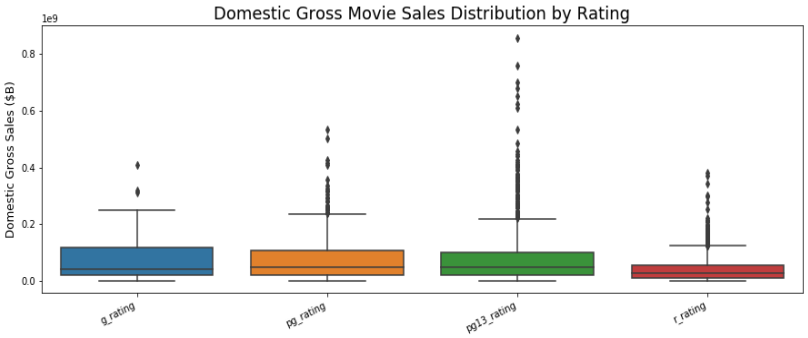
Analysis: The box plots illustrate the distribution of movie release years for 20 years (2000- 2019). The distribution appears to be higher between the years of 2012-2019. Domestic sales are higher after 2012 possibly because the economy was recovering after the Great Recession, which was between 2007-2009.



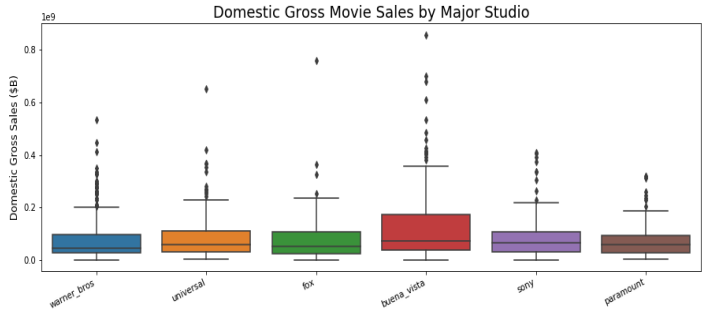
Analysis: The distribution by release week. We see stronger sales performance in the spring and summer months, in line with the traditional summer blockbusters. Additionally, movies released around the November/December holiday time period perform well.



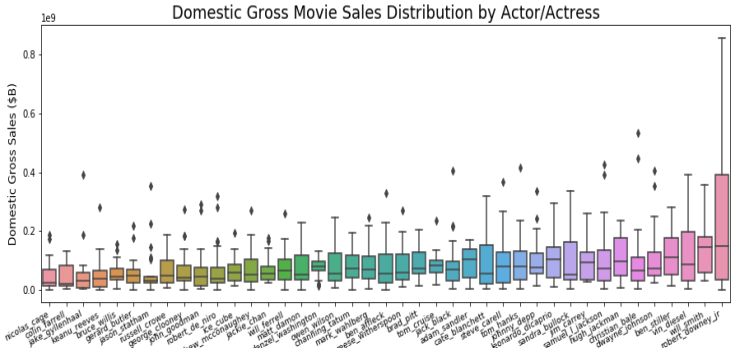
Analysis: Since Robert Downey, Jr. showed the strongest correlation with sales among all actors, we suspected that his role in some very high grossing Marvel Comics films might be the reason. This exhibit shows that RDJR's non-Marvel movies perform better than other films without him and that Marvel films with RDJR perform better than otherwise. This may indicate his presence in a movie is truly a powerful indicator of strong sales (or that he does a good job of selecting good roles).



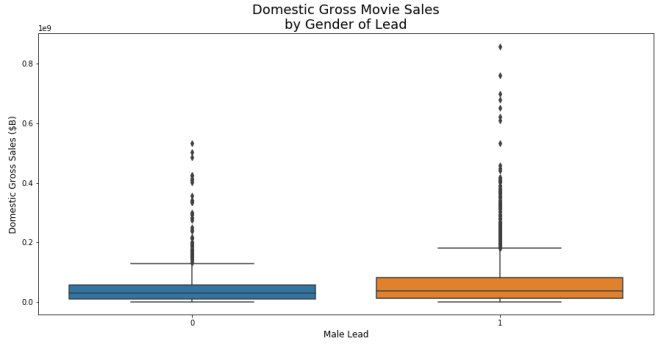
Analysis: Now we look at movie ratings. The pg-13 movie rating scored the highest in sales and r-rating movies ranking the lowest. This distribution intuitively makes sense because wider ranges of individuals are allowed to see a pg-13 rating movie, where as an R-rated movie is limited.



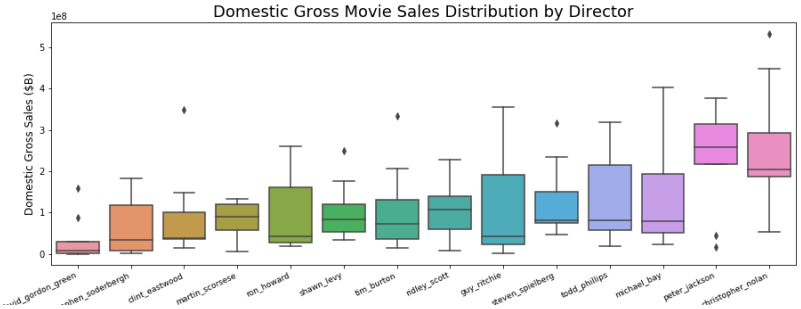
Analysis: When looking at movie studios, it is not surprising that "Buena Vista" (a Disney studio) had the highest movie sales.



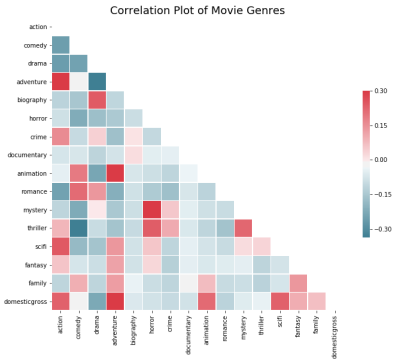
Analysis: When looking at actors and actresses Robert Downey Jr. ranked the highest in garnering 1 billion dollars in domestic sales. This is most likely due to his participation in the Avengers franchise (though note the exhibit above). Vin Diesel was the second actor, probably to due to the Fast and Furious franchise. Sandra bullock was the highest grossing actress.

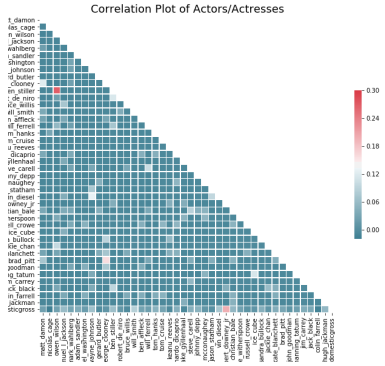


Analysis: When examining gender, male leads grossed the most amount in domestic sales.

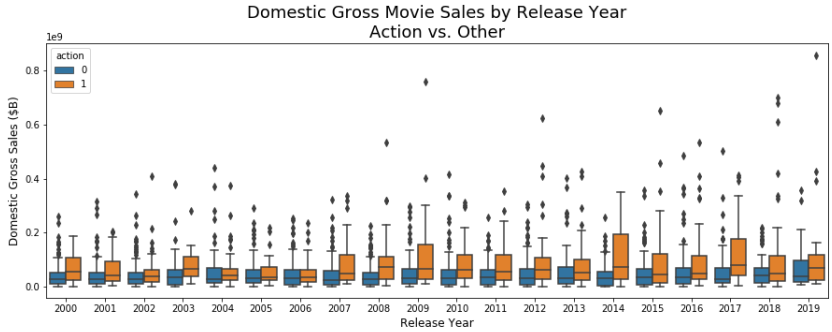


Analysis: When looking at the distribution of directors, Christopher Nolan, had the highest domestic gross, followed by Michael Bay.

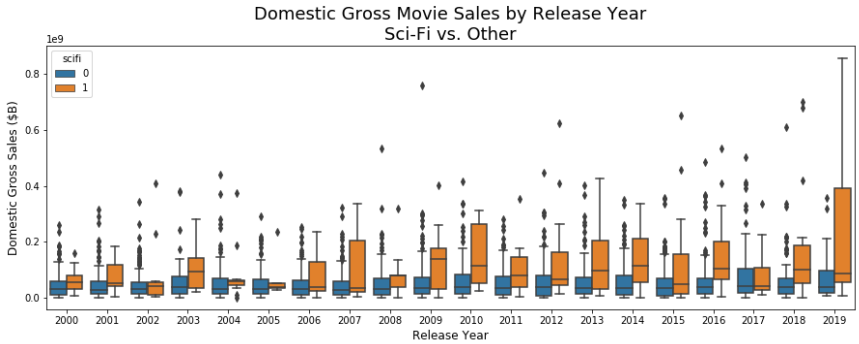




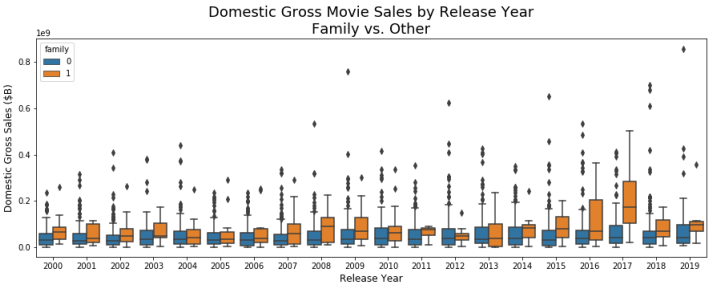
Analysis: In an effort to pick strong features for our model, we tried performing dimensionality reduction for genres. We tried exploring if the different genres (such as family, fantasy, romance etc.) had a relationship with other genres and domestic gross. We experimented with various clustering techniques, but failed to improve model performance through inclusion of clusters or principal components over the raw genre indicators. We see very little correlation among actors/actresses.



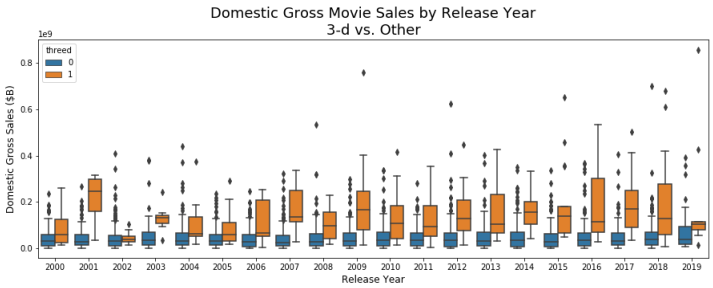
Analysis: These boxplots show the distribution of action movies across domestic movies sales and release years. The blue distribution corresponds to a non-action movie, whereas the orange distribution represents action movies. Across the two decades, action movies, as opposed to non-action movies garnered a higher domestic movie sale. The years 2014 and 2017 had the best performance for action movies.



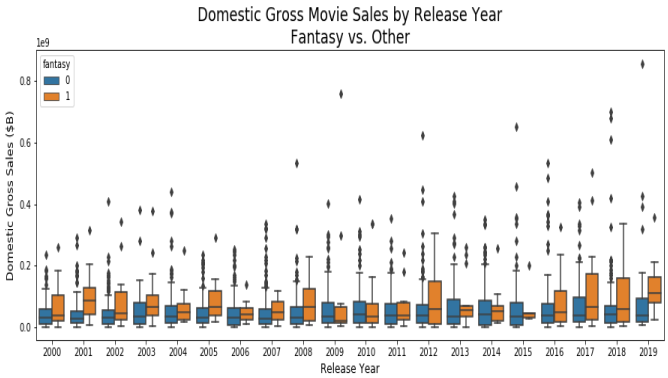
Analysis: This distribution looks at the genre, Sci-Fi across domestic sales and release year. Similar to the action genre in the previous slide, we notice that Sci-Fi also had a strong performance in sales across the two decades. In 2019, we see the highest year in domestic sales for Sci-Fi. This evidently is an up and coming popular genre.



Analysis: This box plot shows the distribution of family films. Family films had a good couple of years such as in 2017 and 2016.



Analysis: This examines the box plot distribution of 3-D movies. There are high sales across the two decades, with a spike in 2016. Typically, movie studios will only invest in 3-D capability for movies they believe will draw high sales, and this affirms that explanation.

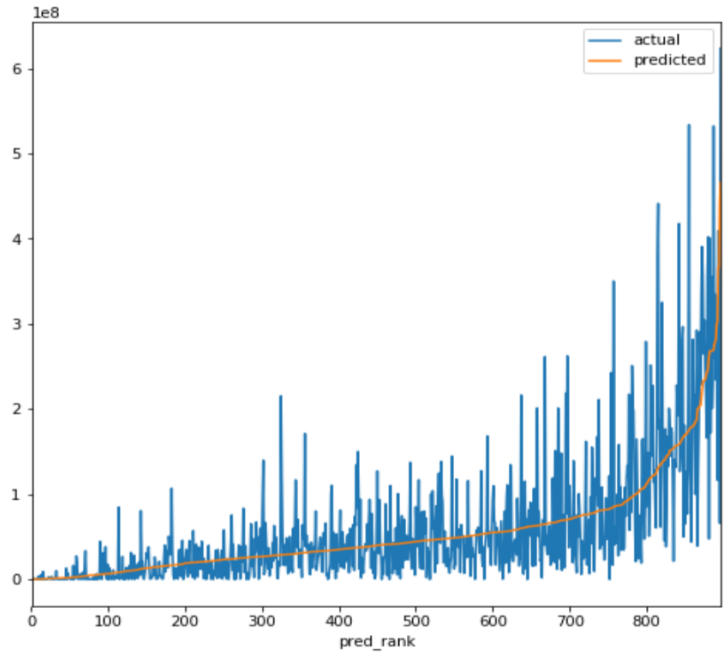


Analysis: We also looked at the distribution of the genre fantasy. This appears to be an emerging genre, as it has outperformed other genres in the last few years.

Build, Tune, and Validate Model

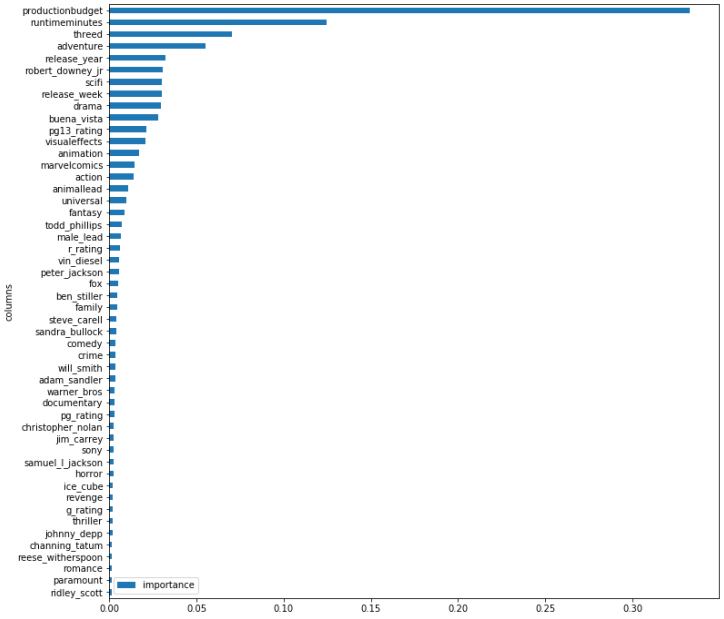
Machine learning (ML) pipelines consist of several steps to train a model. To build better machine learning models, we implemented and tested several machine learning algorithms with the data collected and tuned a subset of the better performing algorithms. After experimenting with some exploratory models, we created a more systematic training pipeline through Scikit-learn. We split the preprocessing steps between categorical and numeric features using *ColumnTransformer*. For the numeric features, we imputed with the median value (only necessary for the runtimeminutes column) and applied a min/max scaler. For the categorical (i.e. one-hot encoded) features, we simply imputed any missing values with 0.

Our initial model evaluation compared Gradient Boosting Regressor, Multi-Layer Perception, Nearest Neighbors Regressor, Bayesian Ridge, Linear Regression, Elastic Net Regression, Lasso Lars and Random Forest Regressor models. After selecting the better performing algorithms, we conducted hyperparameter tuning via Randomized Search and saw improvements in Gradient Boosting Regressor and Random Forest Regressor model performance. In the end, the best performing model was a Gradient Boosting Regressor with an R-squared value of 0.586.

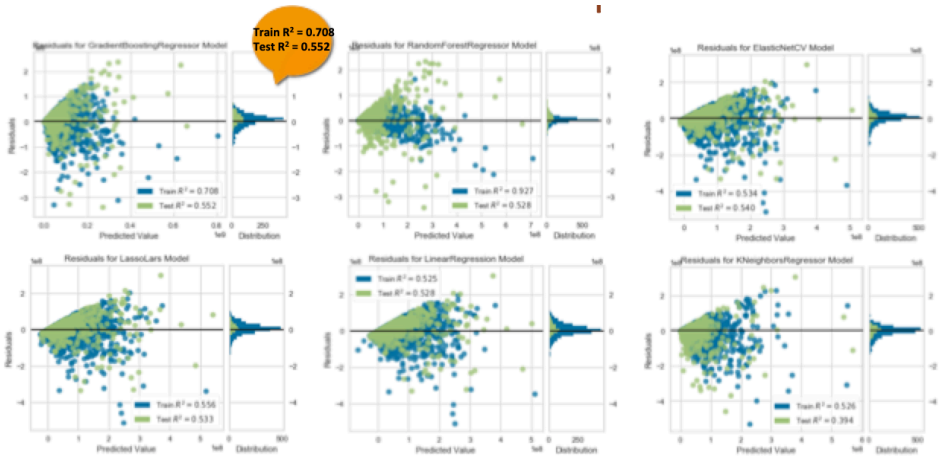


The figure shows the actual versus predicted values on the holdout test dataset.

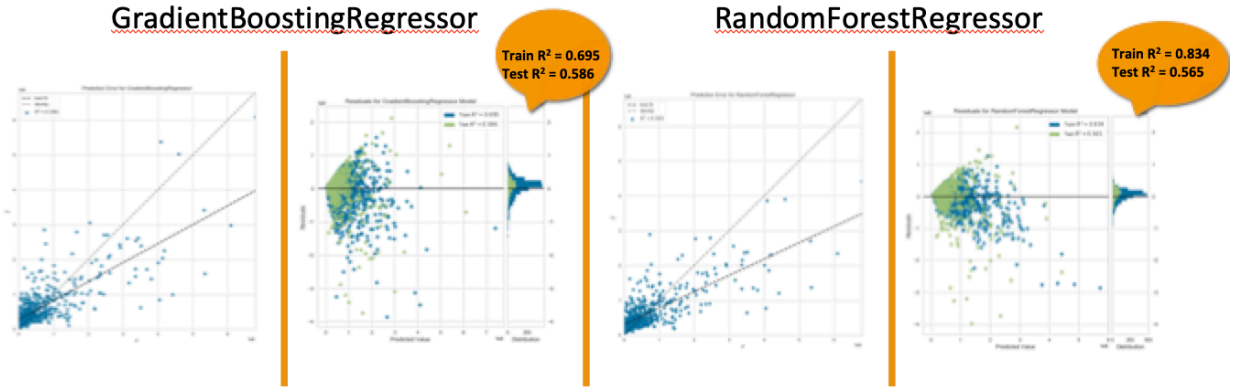
This figure looks at the top 30 variables identified by the model. The most important variables are production budget, run time minutes and if the movie is 3-D.



This figure shows a comparison of the R-squared values for each of the models. The gradient boosting model performed the best.



This figure looks at our best model (gradient) and our random forest model. The tuned Gradient Boosting Regressor outperforms the comparably tuned Random Forest.



Below is a description of the hyperparameters for the final model:

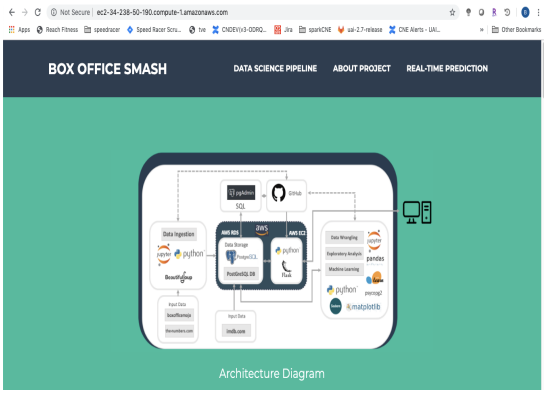
```
GradientBoostingRegressor(alpha=0.9, ccp_alpha=0.0,
criterion='friedman_mse', init=None,
learning_rate=0.05, loss='ls', max_depth=5,
max_features='sqrt', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
```

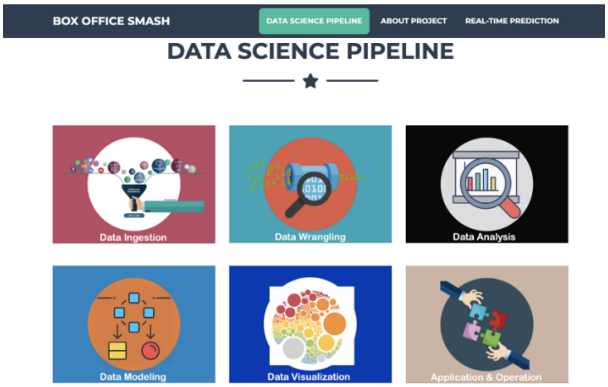
Note that after tuning and selecting our final model, we experimented with some additional feature engineering, including adding features for: Rolling 12-month average sales for primary genre, Average sales for prior films for the lead actor/actress, # prior films for lead actor/actress, Average sales for prior films by the director, and # prior films by the director. Adding these features increased the R-squared value slightly to 0.602, but it potentially complicates the implementation and interpretability of the model.

Real-time Box Office Smash APP prediction

We built a real-time prediction application for our final product on Amazon's EC2 using *Flask*. The application allows users to compute the domestic gross of a movie by inputting the information they know about the movie, such as the cost of production budget, runtime, release year, release week, movie rating, genre, lead actors/actresses, director and the studio. If used widely, movie producers or those in the film industry could get an estimate of the domestic sales accrued by having certain actors/actresses in the film, the number of minutes

the film should be, the director, as well as other features. This information should help those in the film industry allocate their money in the right places.





Link to final product: http://ec2-34-238-50-190.compute-1.amazonaws.com/

Conclusion

Primary challenges for this project included entity resolution in joining datasets from multiple sources and determining the best way to encode genre and actor/actress-related features. With additional effort, we would ideally improve our title match rate and experiment with additional feature engineering, factoring in prior sales figures for movies similar to the movies we're interested in predicting (e.g. looking at performance of sequels).

Additionally, we would attempt to find more datasets describing actor/actress characteristics. By having more information on actors/actresses we could have analyzed the success rate of movies by people who have similar or dissimilar characteristics. Finding more complete datasets would also be helpful for deeper insight on our output.

In all, predictive models for the box office performance of movies were trained by factors derived from Box Office Mojo, IMDb, The Numbers websites data. According to our models, the Gradient Boosting Regressor algorithm significantly improves upon the performance of our baselines. Our R^2 of 0.586 outperformed baseline models and we saw evidence that we could improve it further through additional feature engineering (to 0.602). [5]

We hope our analysis and real-time prediction application will help studios to optimize and position their product in the marketplace.

Github link: https://github.com/georgetown-analytics/Box-Office-Smash/

Sources

[1] Why Solo's Box Office Predictions Were So Wrong

https://screenrant.com/solos-box-office-predictions-wrong/

[2] A Survey on Box-Office Opening Weekend Prediction Using Twitter Data. Engineering and Science SP/2018 vol. 13.

http://www.eurasianjournals.com/A-Survey-on-Box-Office-Opening-Weekend-Prediction-Using-Twitter-Data,109578,0,2.html

[3] Predicting Movie Success from Tweets

http://www.cs.cmu.edu/~rkanjira/files/rose_sneha_ghc17_final.pdf

[4] Early Predictions of Movie Success: the Who, What, and When of Profitability

https://www.biz.uiowa.edu/faculty/kangzhao/pub/JMIS_2016.pdf

[5] Literature Review

https://arxiv.org/abs/1804.03565